IELTS Writing Score Prediction

Introduction

Over 3 million IELTS test are taken every year. The popular exam is prerequisite for admission into many higher education institutions. Additionally, IELTS scores are used as proof of English proficiency by immigration authorities in the U.K., New Zealand, Australia, and Canada.

Grading is done along four major skills: Writing, Reading, Speaking, and Listening. The writing section includes two major components: tasks 1 and 2.

The official IELTS website provides this descripion for task 1:

In IELTS Academic Writing Task 1, you will be shown a diagram, a visual way to represent information. You may be shown one or more than one diagram. This visual information can be shown as a:

- Table
- Chart
- Diagram
- Process
- Graph
- Map

You will also be given the following instructions:
Summarise the information by selecting and reporting the main features and make comparisons where relevant. You will need to do an information transfer task – the visual information you are given needs to be presented in the form of text. As part of the task, you will need to: Write an introduction Write an overview (a summary of what you see) Present and highlight the key features with figures (data) Let's take a closer look at the last three points – the introduction, the overview and the key features that need to be presented.

And the following description for task 2:

IELTS Writing Task 2 requires you to write an essay in response to a statement, or premise. You must read the question carefully so that all parts are answered. For example, in the question below, you must do 3 things to achieve a higher band, showing the examiner that you are addressing all parts of the task.

- Present one view
- Present the other view
- Present your opinion

Grading for both of these sections is conducted by human graders who award or deduct points from the exam-takers based on a detailed rubric linked here. The primary factors considered are task achievement, lexical range, coherence and cohesion, and grammatical range and accuracy. Foundation NLP models are often tasked with determining grammatical accuracy as a demonstration of their language understanding capabilities. In this project, I use BERT, the early encoder model for NLP, to grade both writing tasks. The original BERT paper showcases results on the Corpus of Linguistic Acceptability (CoLA) as a demonstration of the model's ability to learn and understand grammar. With fine-tuning, the BERT model should be able to learn the IELTS rubric, enabling it to grade IELTS writing tasks

Loading the Data

The data is loaded from the csv file ielts_writing_dataset.csv. The dataset is publicly available through Kaggle and can be found here.

Using Pandas, the data is loaded into a Pandas DataFrame, which can be examined.

In [1]:

```
import os
import pandas as pd

data_folder = '/Users/seyedmasihzakavi/Desktop/IELTS_Writing/Data'

df_all = pd.read_csv(os.path.join(data_folder, 'ielts_writing_dataset.csv'))

df_all.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1435 entries, 0 to 1434
Data columns (total 9 columns):
 #   Column             Non-Null Count  Dtype  
---  ------             --------------  -----  
 0   Task_Type          1435 non-null   int64  
 1   Question           1435 non-null   object 
 2   Essay              1435 non-null   object 
 3   Examiner_Commen    62 non-null     object 
 4   Task_Response      0 non-null      float64
 5   Coherence_Cohesion 0 non-null      float64
 6   Lexical_Resource   0 non-null      float64
 7   Range_Accuracy     0 non-null      float64
 8   Overall            1435 non-null   float64
dtypes: float64(5), int64(1), object(3)
memory usage: 101.0+ KB
```

As seen above, there are no null values for Task_Type, Question, Essay, and Overall Score. There are two task types: 1 and 2. Corresponding to each one, we can train a model and examine its power in grading the responses.

To build these two datasets, below, two new Pandas DataFrames are created base on task type, df_task_1 and df_task_2.

In [2]:

```
df_task_1 = df_all[df_all["Task_Type"] == 1][["Question", "Essay", "Overall"]]

df_task_2 = df_all[df_all["Task_Type"] == 2][["Question", "Essay", "Overall"]]
```

Let's consider a few questions and essays for each task to better understand the data

In [4]:

```
pd.set_option('display.max_colwidth', None)

df_task_1.head()
```

Out[4]:

	Question	Essay	Overall
0	The bar chart below describes some changes about the percentage of people were born in Australia and who were born outside Australia living in urban, rural and town between 1995 and 2010.Summarise the information by selecting and reporting the main features and make comparisons where relevant.	Between 1995 and 2010, a study was conducted representing the percentages of people born in Australia, versus people born outside Australia, living in urban, rural, and town. First, in 1995, cities represented the major percentage of habitat by roughly 50 percent, followed by rural areas and towns came in last, among people born in Australia. On the other hand, people born outside Australia, cities showed the most percentages of 6o percent, followed by rural areas and towns. In 2010, among people born in Australia, cities had an increase more than 20 percent increase in the total representation and a major decrease in towns and rural areas. Conversely, people born outside Australia, cities had the most percentage among both studies, followed by rural areas and towns.	5.5
	The bar chart below describes some changes about the percentage of people were born in	The left chart shows the population change happened in Austrilia from 1995 to 2010. In 1995, half of the people born in australia are from cities, 30% from rural areas and only	

	Question	Essay	Overall
2	people were born in Australia and who were born outside Australia living in urban, rural and town between 1995 and 2010.Summarise the information by selecting and reporting the main features and make comparisons where relevant.	20% are from towns. For the people outside of Australia, most of the people still born in cities, which is around 60%. but the number of rural areas increased to 40% with the towns born rate decreased to only 10%.\nIn 2010, The people born in cities increased significianly in both in and outside Australia, especially in outside Australia, which reached 80%. The people bore in towns decreased simutanuously, to around 17% of the people born in Australia and 10% of outside Australia respectively. The most significiant change happened at rual areas numbers. It has shrinked to 17% of people born in Australia, and only around 5% of peopel bore outside Australia.\nOverall, the chart shows us the trend that many people moved to Cities from rual area in the past 15 years.	5.0
4	The graph below shows the number of overseas visitors to three different areas in a European country between 1987 and 2007Summarise the information by selecting and reporting the main features, and make comparisons where relevant.	Information about the thousands of visits from overseas to three different European natural places during 1987 and 2007 is provided in the given line chart.\nOverall, it can be seen that the number of visitors increased significantly in the three places compared to the initial year. Although, visits to Europeans lakes demostrated more changes over the 20 years than its counterparts.\nIn more detail, the most steady growth was experienced by the visits to Europeans mountains. For example, from 1987 the number of visitors grew from 20,000 to almost the double 20 years later. Similarly, visits to the coast also rose after a slight fall in 1992, reaching almost twice as much since 1987, with 75,000.\nThose visiting Europeans lakes subtantially increased over the years from 10 thousand to a peak of 75 thousand in 2002. Despite falling for about 25 thousand in 2007, the visitis to this place remained higher compared to 1987, with 50,000 at the end of the period.	7.0
6	The graph below shows the number of overseas visitors to three different areas in a European country between 1987 and 2007Summarise the information by selecting and reporting the main features, and make comparisons where relevant.	This graph depicts the changes in tourists visits between 1987 and 2007 to these three places in Europe. It's clear from the graph, that generally the coast area have the highest number of visitors all over the years.\nAccording to what is shown in the linear figure in 1987 period visitors first place was the coast with 40 visitor and in seconed place was the mountains with half number of coast's visitors and the lakes became lastly.\nWhile the statistics have been changed at 2002 , the lakes sharply increased to be the first place by reaching the highest number of it's visitors, and the coast is the seconed.\nFinally at 2007 , the coast speedly increased to be most place having tourists and the lakes dramaticly declined to take the seconed place. The rules switched between coast and lakes numbers.\nTo sum up , the linear graph describe a constantly change in mountains visitors numbers , and sharply changes in coast and lakes visitors numbers.	6.5
8	The graph below shows the number of overseas visitors to three different areas in a European country between 1987 and 2007Summarise the information by selecting and reporting the main features, and make comparisons where relevant.	The line graph illustrates the number of overseas tourists to three different areas in a European country (coast, mountains, lakes) from 1987 to 2007\nOverall, it can be seen that the number of visitors has increased for the three categories of areas over the period given. Furthermore, in 2007 the highest number of visitors was reported on the coast at approximately 85 thousand. The second most popular areas were the lakes and last were the mountains.\nIn regard to the coast, we can see that the number of visitors from 1987 to 1992 decreased. However, from 1992 to 2007 there was a consistently increased number of tourists climbing to over 70 thousand visitors in the final year of analysis.\nAbout the mountains, the number of overseas tourists begat at 20 thousand and has increased to reach over 30 thousand by 2007.\nFinally, the lakes were visited by 10 thousand tourists in 1987 and rose significantly to reach over 70 thousand in 2002 and then decreased to 50 thousand by 2007.	8.0

Note that the section 1 prompt includes graphs and charts. Here, those are omitted to make training feasible on a smaller NLP model that can be trained on a local machine with limited computational resources

In [5]:

```
df_task_2.head()
```

Out[5]:

	Question	Essay	Overall
1	Rich countries often give money to poorer countries, but it does not solve poverty. Therefore, developed countries should give other types of help to the poor countries	Poverty represents a worldwide crisis. It is the ugliest epidemic in a region, which could infect countries in the most debilitating ways. To tackle this issue, rich countries need to help those in need and give a hand when possible. I agree that there are several ways of aiding poor countries other than financial aid, like providing countries in need with engineers, workers, and soldiers who would build infrastructure. Building universities, hospitals, and roadways. By having a solid infrastructure, poor countries would be able to monetise their profits and build a stronger and more profitable economy which would help them in the long term. Once unprivilged countries find their niche, the major hurdle would be passed and would definitely pave the way for much brighter future. However, I do disagree that financial aid does not solve poverty, it does if used properly and efficiently. The most determining factor if financial aid would be the way to go, is by identifying what type of poor countries' representative are dealing with. Some countries will have a responsible leader and some will not, with that being said, implementing a strategy, to distinguish responsible leaders from others, would tailor the type of aid rich countries could use. An example, A clear report and constant observation would be	6.5

	Question	Essay	Overall
	rather than financial aid. To what extent do you agree or disagree?	applied to track the progress and how this type of aid is being monetized. In summary, types of aid varies from country to another, and tailoring the type of aid is of paramount importance to solve this problem that had huge toll on poor countries.	
3	Rich countries often give money to poorer countries, but it does not solve poverty. Therefore, developed countries should give other types of help to the poor countries rather than financial aid. To what extent do you agree or disagree?	Human beings are facing many challenges nowadays. Poverty is always an critical topic among countries, especially the poverty in developing countries. Developed countries frequently offer financial support to poor countries but the poverty still exist. Experts are arguing that developed countries should consider other solutions to help solve the poverty issue. I believe this is a much better direction compare to money support only.\nMoney is essential to many factors like food, contruction and hospital. With money provided by developed countries, govenment can improve the inforstructure, supporting poor family with food, building more houses for the poor people, builing more schools to support children's education, which is critical for people to get out of poverty.\nBut money cannot solve everything. The poverty caused by many reasons like lacking of resource and experienced governor. None of them could be solved by simply offering money to the goverment. Besides, money could cause a bad habit to the governor, they may reply on the financial support too much. If the financial support ended someday, which could happen anytime in the current circurmestance, it will cause big trouble again.\nDevevloped countries should focus on a sophicated solution instead of just providing money. For example, they can help poor countries to build manufacturing industry, and give certain subsides for the customer who buy directly from them. Developed countries also can send experienced instructor to help governer in poor countries build a better government system to improve the efficiency of governance, authority of justice, to make sure company feel safe to invest in these developing counties, which could have a long term benifit for the people there.\nOverall, I am agree that money could help to solve the poverty but without other support from different functions, it may not the best solution for poverty.	5.5
5	Some countries achieve international sports by building specialised facilities to train top athletes, instead of providing sports facilities that everyone can use. Do you think this is positive or negative development? Discuss both views and give your opinion.	Whether countries should only invest facilities and training on their elite athletes in order to win international competitions or give the same opportunities for all remains a big discussion. In this essay I will explain both views and why I think participation for all is the way which brings positive consequences.\nThose supporting the first view believe that the odds to win are only high if the countries invest specialised facilities on their top sportmen. This is because, only athletes trained to reach the highest performance guarantee high scores. For example, it is known that China sends only its elite sportmen to compete in the Olympics and statistically, they usually win numbers of medals.\nOn the other hand, those who support sports facilities and training for all athletes consider that non-top ones can also have great opportunities to succeed. For instance, Diego Maradona, a Argentinian soccer player, never played in a professional field, on the contrary, he was discovered playing in an old field located in a poor village. This means that by providing facilities for all, gifted sportmen might be discovered and they can make history too.\nIn my opinion, as a supporter of the second view, I believe that permiting everyone to train and use sports facilities is the best way to achieve international sports. For me, winning is as essential as participation. Besides, this way leads to find natural talents among the local population, which is inspiring for others in similar circunstances.\nIn conclusion, although the two positions have strong arguments. I think that the one supporting facilities for all people is the one that can bring positive developments to achieve sports events. Participation and inclusiveness may help to find gifted athletes among ordinary people.	6.5
7	Some countries achieve international sports by building specialised facilities to train top athletes, instead of providing sports facilities that everyone can use. Do you think this is positive or negative development? Discuss both views and give your	Sports is an essential part to most of us , some of us consider it a lifestyle and others consider it a job or source of income , So it's a huge part in our live.\nBuilding specialized facilities to improve and develop athelets levels it's kind of encouragment to this category of people , and it reflects enormous benefit to them. Either by providing to them the professional training or by make them ready to international competitions. But narrowing the attention of sports to a specific category and ignore other categories actually doesn't sounds a good idea.\nAthelets consider as a important part of the community but the don't cover most of it, and other people have the right to have sport facilities that suits their level and everyone whatever is it beginner or professional can use it.\nSports is a public interest that doesn't include specific people in it, and it's important to all genders and all ages, so providing facilities to specific ones can end up with unfairness results. In other hand , providing this facilities to all public people will lead to increase in sport awarness and end up with more people having healthy lifesyle.\nIn conclousion , according to medical facts said that's doing sport daily and make it essence in your life decreases heart attacks and raise the life quality, specified just the athelets people with sports facilities is undesireable decision. In other hand , raising the number of gyms and sports areas will conclude with healthier society.	5.5

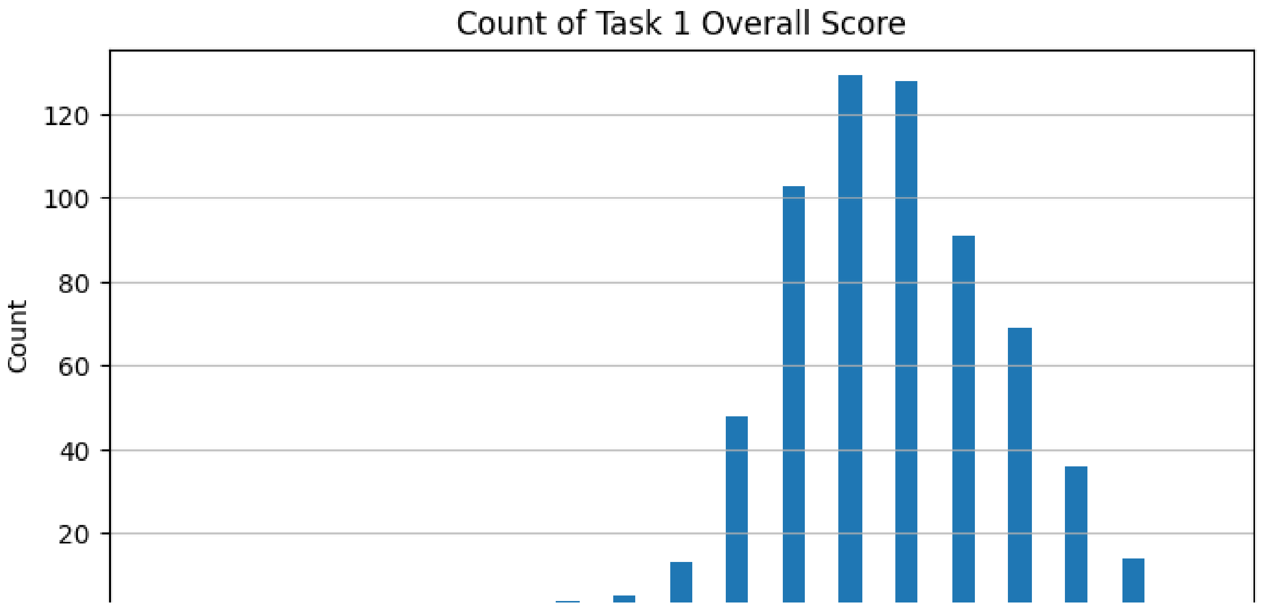
	opinion. Question	Essay	Overall
9	Some countries achieve international sports by building specialised facilities to train top athletes, instead of providing sports facilities that everyone can use. Do you think this is positive or negative development? Discuss both views and give your opinion.	International sports events require the most well-trained athletes for each country, in order to achieve this goal countries make an effort to build infrastructure designed to train top athletes. Although this policy can indeed make fewer sports facilities for ordinary people, investing in the best athletes is vital to develop competitive sports performances in each country.\nOn the one hand, building specific infrastructure for the best athletes is crucial in order to get better results at international sports events such as The Olympics or the World Cup. The importance of getting better results is that it creates awareness of the importance of sports in society and motivates more people to do a sport. In this way, investing in these developments can help countries to develop an integral sport policy that can benefit everyone.\nOn the other hand, one can argue that a negative effect could be that less infrastructure is built for the rest of the people. However, people who practice a sport in their daily life do not necessarily need some facilities to do sports. For example, people often use public spaces to do sports such as running or doing yoga at the nearest park to their home. So, for people who is not top athletes there could be some alternatives for sports facility that ,is not the case for training top athletes.\nTo sum up, I strongly believe countries should invest in specialised infrastructure for their best athletes because in the long term is going to generate more motivation to do sports, to invest in sports at schools and therefore to build more sports infrastructure for everyone.	9.0

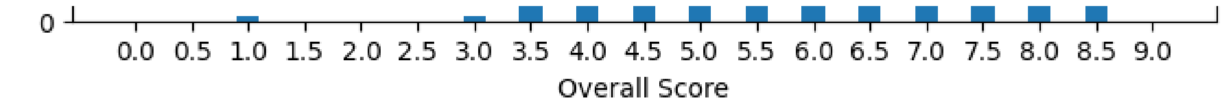
Exploratory Data Analysis

To better understand the data, we can visualize the distribution of scores for each task

In [6]:

```
import matplotlib.pyplot as plt

scores = [i*0.5 for i in range(19)]
task_1_scores = [0 for _ in range(len(scores))]
task_2_scores = [0 for _ in range(len(scores))]

for i in range(len(scores)):
    task_1_scores[i] = len(df_task_1[df_task_1['Overall'] == scores[i]])
    task_2_scores[i] = len(df_task_2[df_task_2['Overall'] == scores[i]])

plt.figure(figsize=(8, 4))
plt.bar(scores, task_1_scores, width=0.2)
plt.title('Count of Task 1 Overall Score')
plt.xlabel('Overall Score')
plt.ylabel('Count')
plt.xticks(scores)
plt.grid(axis='y', alpha=0.75)
plt.show()
```

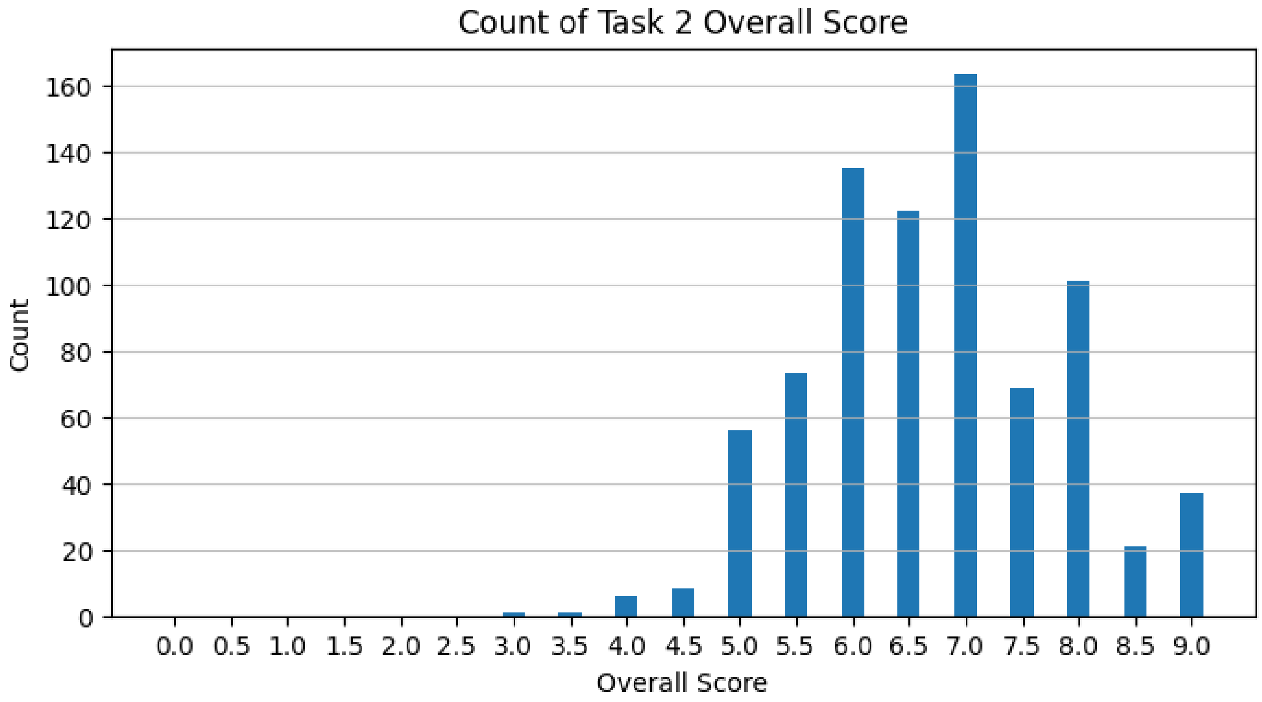




Notice that scores have a normal distirbution centered around 6 - 6.5. Different scores are very unevenly distributed for Task 1, with almost no entries with scores lower than 4 or 9. Below, the same bar plot can be seen for Task 2.

In [7]:

```
plt.figure(figsize=(8, 4))
plt.bar(scores, task_2_scores, width=0.2)
plt.title('Count of Task 2 Overall Score')
plt.xlabel('Overall Score')
plt.ylabel('Count')
plt.xticks(scores)
plt.grid(axis='y', alpha=0.75)
plt.show()
```



Similar to Task 1, Task 2 scores are normally distributed and centered around 6 - 7.

Task 1 Model

The model is built on top of BERT to leverage pre-trained weights. The base BERT model from the original paper with 12 layers and attention heads is used here.

In [8]:

```
from transformers import BertForSequenceClassification, BertConfig, AutoTokenizer

config = BertConfig.from_pretrained('bert-base-uncased',
                                    num_labels=1,
                                    problem_type="regression")


model_1 = BertForSequenceClassification.from_pretrained('bert-base-uncased',
                                                        config=config)
```

```
/Users/seyedmasihzakavi/Library/Python/3.9/lib/python/site-packages/urllib3/__init__.py:3
5: NotOpenSSLWarning: urllib3 v2 only supports OpenSSL 1.1.1+, currently the 'ssl' module
is compiled with 'LibreSSL 2.8.3'. See: https://github.com/urllib3/urllib3/issues/3020
  warnings.warn(
/Users/seyedmasihzakavi/Library/Python/3.9/lib/python/site-packages/tqdm/auto.py:21: Tqdm
```

```
Warning: IProgress not found. Please update jupyter and ipywidgets. See https://ipywidget
s.readthedocs.io/en/stable/user_install.html
  from .autonotebook import tqdm as notebook_tqdm
Some weights of BertForSequenceClassification were not initialized from the model checkpo
int at bert-base-uncased and are newly initialized: ['classifier.bias', 'classifier.weigh
t']
You should probably TRAIN this model on a down-stream task to be able to use it for predi
ctions and inference.
```

While only the responses are graded, the questions also contain crucial information which affects the grading. This means both the question and answer need to be provided as input to the model.

The original BERT paper linked here includes fine-tuning results for question answering tasks. While grading IELTS writing tasks is a regression task, it is very close to question answering tasks from the original paper like SQuAD so it makes the most sense to feed the question and asnwers to the model in the same way. As seen in the screenshot below from the original paper, questions and answers are concatenated and then passed to the model.

Following this design, questions and answers will be concatenated with a special [SEP] token in between.

In the section below, three separate DataFrames are built: Training, Validation, and Testing with 80%-10%-10% ratios. Then the question and answer are concatenated with a [SEP] token and stored in a new column, combined_text. Finally, the BERT tokenizer needs to be applied to the concatenated text before it's fed into the BERT model. This step happens in the EssayDataset class, which implements PyTorch's Dataset and applies the tokenizer before accessing each combined question and essay.

In [9]:

```
from sklearn.model_selection import train_test_split
from torch.utils.data import Dataset
from transformers import AutoTokenizer
import torch


task_1_train, task_1_temp = train_test_split(df_task_1, test_size=0.2, random_state=42)
task_1_val, task_1_test = train_test_split(task_1_temp, test_size=0.5, random_state=42)


tokenizer = AutoTokenizer.from_pretrained('bert-base-uncased')

def prepare_dataset(df):
    df['combined_text'] = df['Question'] + " [SEP] " + df['Essay']
    combined_texts = df['combined_text'].tolist()
    labels = df['Overall'].tolist()

    tokenized_batch = tokenizer(combined_texts, padding="max_length",
                                truncation=True, max_length=512)

    class EssayDataset(Dataset):
        def __init__(self, encodings, labels):
            self.encodings = encodings
            self.labels = labels

        def __getitem__(self, idx):
            item = {key: torch.tensor(val[idx]) for key, val in self.encodings.items()}
            item['labels'] = torch.tensor(self.labels[idx], dtype=torch.float)
            return item

        def __len__(self):
            return len(self.labels)

    return EssayDataset(tokenized_batch, labels)

task_1_train_dataset = prepare_dataset(task_1_train)
task_1_val_dataset = prepare_dataset(task_1_val)
task_1_test_dataset = prepare_dataset(task_1_test)
```

BERT's architecture and weights are downloaded through Hugging Face's Transformers library. The model is set to train for 10 epochs, go through linear warmup for 500 steps, and save weights for the model with lowest mean squared error on validation data.

In [10]:

```python
import torch
from transformers import BertForSequenceClassification, TrainingArguments, Trainer
from sklearn.metrics import mean_squared_error
import numpy as np


training_args_1 = TrainingArguments(
    output_dir='./results_1',
    num_train_epochs=10,
    per_device_train_batch_size=8,
    per_device_eval_batch_size=8,
    warmup_steps=500,
    evaluation_strategy="epoch",
    save_strategy="epoch",
    learning_rate=1e-5,
    load_best_model_at_end=True,
    metric_for_best_model="mse",
    greater_is_better=False
)

def compute_metrics(eval_pred):
    predictions, labels = eval_pred
    mse = mean_squared_error(labels, predictions)
    return {"mse": mse}

trainer_1 = Trainer(
    model=model_1,
    args=training_args_1,
    train_dataset=task_1_train_dataset,
    eval_dataset=task_1_val_dataset,
    compute_metrics=compute_metrics
)
```

```
Some weights of BertForSequenceClassification were not initialized from the model checkpo
int at bert-base-uncased and are newly initialized: ['classifier.bias', 'classifier.weigh
t']
You should probably TRAIN this model on a down-stream task to be able to use it for predi
ctions and inference.
/Users/seyedmasihzakavi/Library/Python/3.9/lib/python/site-packages/accelerate/accelerato
r.py:432: FutureWarning: Passing the following arguments to `Accelerator` is deprecated a
nd will be removed in version 1.0 of Accelerate: dict_keys(['dispatch_batches', 'split_ba
tches', 'even_batches', 'use_seedable_sampler']). Please pass an `accelerate.DataLoaderCo
nfiguration` instead:
dataloader_config = DataLoaderConfiguration(dispatch_batches=None, split_batches=False, e
ven_batches=True, use_seedable_sampler=True)
  warnings.warn(
```

In [11]:

```python
trainer_1.train()

model_path = "./results_1"
model_1.save_pretrained(model_path)
tokenizer.save_pretrained(model_path)
```

```
  0%|          | 0/650 [00:00<?, ?it/s]huggingface/tokenizers: The current process just g
ot forked, after parallelism has already been used. Disabling parallelism to avoid deadlo
cks...
To disable this warning, you can either:
 - Avoid using `tokenizers` before the fork if possible
 - Explicitly set the environment variable TOKENIZERS_PARALLELISM=(true | false)

 10%|█         | 65/650 [01:53<12:57,  1.33s/it]
```

```
{'eval_loss': 34.44484329223633, 'eval_mse': 34.444847106933594, 'eval_runtime': 3.8947,
'eval_samples_per_second': 16.433, 'eval_steps_per_second': 2.054, 'epoch': 1.0}
```

```
 20%|██        | 130/650 [03:49<11:08,  1.29s/it]
```

```
{'eval_loss': 22.51834487915039, 'eval_mse': 22.518342971801758, 'eval_runtime': 3.8045,
'eval_samples_per_second': 16.822, 'eval_steps_per_second': 2.103, 'epoch': 2.0}
```

```
 30%|███       | 195/650 [05:43<09:50,  1.30s/it]
```

```
{'eval_loss': 6.526093006134033, 'eval_mse': 6.526093482971191, 'eval_runtime': 3.7795, '
eval_samples_per_second': 16.933, 'eval_steps_per_second': 2.117, 'epoch': 3.0}
```

```
 40%|████      | 260/650 [07:46<09:03,  1.39s/it]
```

```
{'eval_loss': 0.9443441033363342, 'eval_mse': 0.9443441033363342, 'eval_runtime': 4.1961,
'eval_samples_per_second': 15.252, 'eval_steps_per_second': 1.907, 'epoch': 4.0}
```

```
 50%|█████     | 325/650 [09:50<07:34,  1.40s/it]
```

```
{'eval_loss': 0.8789247274398804, 'eval_mse': 0.8789247870445251, 'eval_runtime': 4.2821,
'eval_samples_per_second': 14.946, 'eval_steps_per_second': 1.868, 'epoch': 5.0}
```

```
 60%|██████    | 390/650 [11:57<06:03,  1.40s/it]
```

```
{'eval_loss': 0.5735933780670166, 'eval_mse': 0.5735933780670166, 'eval_runtime': 4.214,
'eval_samples_per_second': 15.187, 'eval_steps_per_second': 1.898, 'epoch': 6.0}
```

```
 70%|███████   | 455/650 [14:05<04:35,  1.41s/it]
```

```
{'eval_loss': 0.6851528882980347, 'eval_mse': 0.6851528882980347, 'eval_runtime': 4.3627,
'eval_samples_per_second': 14.67, 'eval_steps_per_second': 1.834, 'epoch': 7.0}
```

```
 77%|███████▋  | 500/650 [15:32<04:42,  1.88s/it]
```

```
{'loss': 11.8296, 'grad_norm': 16.327892303466797, 'learning_rate': 1e-05, 'epoch': 7.69}
```

```
 80%|████████  | 520/650 [16:13<03:10,  1.47s/it]
```

```
{'eval_loss': 0.7569351196289062, 'eval_mse': 0.7569351196289062, 'eval_runtime': 4.4528,
'eval_samples_per_second': 14.373, 'eval_steps_per_second': 1.797, 'epoch': 8.0}
```

```
 90%|█████████ | 585/650 [18:37<01:57,  1.81s/it]
```

```
{'eval_loss': 0.6349007487297058, 'eval_mse': 0.634900689125061, 'eval_runtime': 5.2919,
'eval_samples_per_second': 12.094, 'eval_steps_per_second': 1.512, 'epoch': 9.0}
```

```
100%|██████████| 650/650 [21:05<00:00,  1.59s/it]
```

```
{'eval_loss': 0.6780698299407959, 'eval_mse': 0.6780698895454407, 'eval_runtime': 4.7661,
'eval_samples_per_second': 13.428, 'eval_steps_per_second': 1.679, 'epoch': 10.0}
```

```
100%|██████████| 650/650 [21:08<00:00,  1.95s/it]
```

```
{'train_runtime': 1268.0904, 'train_samples_per_second': 4.045, 'train_steps_per_second':
0.513, 'train_loss': 9.182725119957556, 'epoch': 10.0}
```

Out[11]:

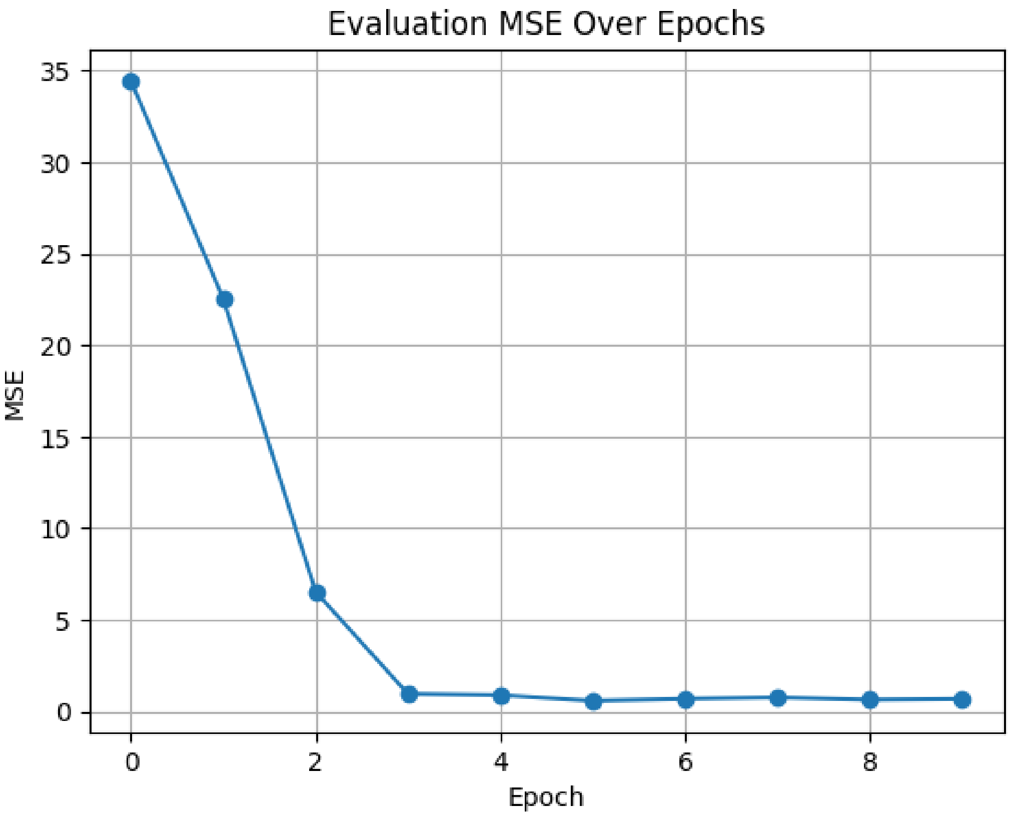
```
('./results_1/tokenizer_config.json',
 './results_1/special_tokens_map.json',
 './results_1/vocab.txt',
 './results_1/added_tokens.json',
 './results_1/tokenizer.json')
```

Using the saved trainer_state file, we can visualize the MSE loss per each epoch, which decreases sharply indicating that the model has indeed learned the patterns in grading the essays and can achieve scores similar

to those assigned by the IELTS Writing graders

In [12]:

```
import json

with open('./results_1/checkpoint-650/trainer_state.json', 'r') as file:
    trainer_state = json.load(file)

mse_values = [log['eval_mse'] for log in trainer_state['log_history'] if 'eval_mse' in l
og]

plt.plot(mse_values, marker='o', linestyle='-')
plt.title('Evaluation MSE Over Epochs')
plt.xlabel('Epoch')
plt.ylabel('MSE')
plt.grid(True)
plt.show()
```



Now that the training is completed, we can evaluate the model's performance on the test data frame.

In [13]:

```
task_1_test_results = trainer_1.evaluate(task_1_test_dataset)
print(task_1_test_results)
```

```
100%|██████████| 9/9 [00:03<00:00,  2.57it/s]
```

```
{'eval_loss': 0.635013997554779, 'eval_mse': 0.6350140571594238, 'eval_runtime': 4.4422,
'eval_samples_per_second': 14.632, 'eval_steps_per_second': 2.026, 'epoch': 10.0}
```

The model achieves an impressive MSE of 0.635 on the unseen test dataframe. To better understand this, let's consider 10 actual essays from the test dataset together with their predicted and actual scores

In [15]:

```
predictions_output = trainer_1.predict(task_1_test_dataset)
predictions = predictions_output.predictions

predictions = predictions.squeeze()
```

```
rounded_predictions = 0.5 * np.round(predictions / 0.5)
task_1_test['predicted_score'] = rounded_predictions
sample_df = task_1_test.sample(n=10, random_state=42)

sample_df[['Question', 'Essay', 'Overall', 'predicted_score']]
```

```
100%|██████████| 9/9 [00:03<00:00,  2.56it/s]
```

Out[15]:

	Question	Essay	Overall	predicted_score
1034	The diagram below shows the life cycle of a salmon, from egg to adult fish.Summarise the information by selecting and reporting the main features and make comparisons where relevant.	The diagram illustrates the lifecycle of a salmon from spawning phase to maturity. The life of a salmon starts in the incubation period inside the eggs which is spawned in the freshwater. After the incubation period the salmon emerge from the eggs in the river and they are reared in the freshwater. The young salmon moves through the freshwater in the rives towards the estuary and completes the estuary rearing there before they migrate to the ocean.\nOnce they are in the ocean the salmon moves around searching for rearing areas and reaches its full growth and maturity. The matured salmon then migrates to spawning areas in the freshwater through the estuary for spawning. Once the salmon is emerged it goes through several stages of life to reach the maturation phase\nOverall the life cycle illustration of a salmon goes through 8 stages which is started by spawning stage of a salmon and ends when a salmon attains full growth and maturity.	5.5	7.0
352	The charts below show the percentage of Australian men and women in three age groups who were employed in 1984, 2001 and 2014.Summarise the information by selecting and reporting the main features, and make comparisons where relevant.	These bar graphs shows who were employed among Australian females and males in three different years 1984, 2001 and 2014 including the age group of 15-19, 35-44 and 60-64.\nIt is clear from the graph that the age group 35-44 years old has the highest rate of employment among the three years of 1984, 2001 and 2014 and in both sexes.\nAccording to what is shown, in men bar graph, specifically in the 35-44 age group, the percentage did not change through out the years. While the younger group 15-19 employment percentage has been increasing from 1984 to 2001 and 2014. Additionally, older Australian people category 60-64 years old who were employed in 2014 is more than 1984 and 2001.\nOn the other hand, women in the middle age group 35-44 employment got higher with time from the year 1984 to 2001 and 2014. Furthermore, teenagers from 15-19 agre group slightly decreased in employment level in 2014. While elderly from 60-64 got more jobs in 2014.\nIn conclusion, we could say that the two age groups of 35-44 and 60-64 had higher percentage in 2014. In contrast, age group 15-19 had lower percentage by the year 2014.	6.5	6.5
886	The table below shows the salaries of secondary/high school teachers in 2009. Summarize the information by selecting and reporting the main features and make comparison where relevant.You should write at least 150 words.	The table indicates the annual income of secondary/ high school teachers and comparisons in 2009. Overall, it can be seen an increase from starting year and after 15 years in all countries.\nTo go into detail, there are two countries, in Australia, and Denmark where teachers' annual income reached the maximum salary within 10 years. Specifically, The annual income started at 28,000, 45,000 in Australia, and Denmark respectively. Afterwards, the salaries reached 48,000, 54,000 within 9 years, and 8 years in Australia and Denmark.\nWhen it comes to other three countries, Luxembourg, Japan, and Korea, those three countries teachers steadily increased their annual income, however, they kept working after 15 years and reach maximum later. In Japan, and Korea, their annual income gradually increase the longer they worked. For example, the first years' annual income in each country is 34,000, 30,000 in Japan, and Korea respectively. After 15 years, their incomes soared to 65,000 and 48,000. As a result, their maximum income will reach the peak of 37 years, and 34 years in Japan, and Korea. The salary is 86,000, 64,000 in Japan, and Korea.\nInterestingly, in Luxembourg, the annual saralies start at 80,000 and increased to 119,000 after 15 years. In contract, there is not so huge gap between after 15 years annual income and maximum. It is 119,000 and 132,000. Furthermore, the maximum salary tend to reach in 30 years. It is a litte bit earlier than Japan, and Korea.	7.5	6.5
426	The charts shows air pollution levels by different causes among four countries in 2021.Summaries the information by selecting and reporting the main features, and make	The bar chart illustrates the level of air pollution caused by three different sectors namely electricity generation, transport and industry among four countries which are China, America, Japan and Australia in 2021.\nOverall, it can be seen that China produced the highest amount of the pollution in every sector. Another noticeable point is that all causes in Australia created the pollution at almost the same amount.\nLooking at the detail, the industry, the transportation and the electricity generation in China generated the pollution level of 80, 70 and a little less than 60 ppm respectively. This pattern also occurred in Japan where the industry ranked first in releasing the pollution (above 40 ppm) and the power generation ranked last (produced around 15 ppm).\nFurthermore, approximately 60 ppm of the pollution in Australia is arisen from the electricity generation and the industry. This generated	6.0	7.0

	Question	Essay	Overall	predicted_score
	comparisons where relevant.	pollution of each aforementioned factors accounted for about 30 ppm which is nearly two times less than that of the transport. Focusing further on Australia, all three businesses created the pollution of just below 20 ppm.		
1102	The line graph below shows the number of annual visits to Australia by overseas residents. The table below gives information on the country of origin where the visitors came from.Write a report for a university lecturer describing the information given.	The table and the graph illustrate information about the number of visitors of Australia in million from foreign countries from 1975 to 2005.\nOverall, the number of overseas residents who came to Australia incresed in a straight-line progression from 8.8 to 30.4 during mentioned period of time. The most visitors was from Japan - 3.2 and 12.0, whereas the least number was from China - 0.3 and 0.8 - in 1975 and 2005, respectively.\nPeople from South Korea were in the second place based on the number of visitors - 2.9 in 1975 and 9.1 in 2005. In the third place there was European nations - 1.1 and 4.5 at the same time. The number of Britans was 0.9 and 2.9, whereas the number of americans was 0.4 and 1.1, in 1975 and 2005, correspondingly.\nIt worth to mentioned that that the whole number of foreign visitors of Australia in 1985 and 1995 was approximately 17 and 25, accordingly.	6.5	6.5
992	The chart below shows the results of a survey of people who visited four types of tourist attraction in Britain in 1999.Summarise the information by selecting and reporting the main features and make comparisons where relevant.	The graph presents the information about the proportion of people who visited several different kinds of tourist spots in 1999 in Britain.\nThe persentage of tourists who traveled to the theme parks was the highest at 38%. The figure for museums & galleries was lower at 37%. 16% of visitors chose the historic houses & monuments as the destination. While the wildlife parks &zoos were the most unpopular tourist attractions, only 9% of people chose it.\nAs the most popular place, theme parks have various types that provide for the public. The proportion of visitors who travel the blackpool pleasure beach was the highest at 47%. The figure for alton towers and pleasureland, southport remain the same level at about 16%. Chessington world of adventures and legoland, windsor show the same partten, both 10%.\nOverall, Most of people liked to vist theme parks including blackpool pleasure beach and those kind of things.	7.0	6.5
216	The pie charts below show how dangerous waste products are dealt with in three countries.Write a report for a university, lecturer describing the information shown below.	The three pie graphs provide information on which methods is used to deal with hazardous waste products by Republic of Korea, Sweden and the UK.\nOverall, the main stricking feature is that every countries has a primary way to engage with dangerous end-of-life materials, despite using equally all of them.\nThe specialisation is particulary clear in the UK where the waste burryed underground cover 82% of the total, whereas in both Republic of korea and Sweden the main way of dealing with it is less then three-quarters of the total. Although for the UK and Sweden the method used is dump the materials underground, Republic of Korea favorites recycling, a process that is not used in the UK and covers only 25% in Sweden.\nIt is also evident that Republic of Korea and Sweden proceed in two secondary ways, which sees incineration as a last resources (9% and 20% respectively), while the 18% of the UK uncoverd by the undergorund method is split evenly between chemical treatment and dumping at sea (processes not empolyed by the other nations) and only a mere 2% get destroyed by fire.	6.0	7.0
1200	The table below gives information about the underground railway systems in six cities.Summarise the information by selecting and reporting the main features, and make comparisons where relevant.	The table illustrates information about the underground railway system in six significant cities around the world such as (London, Paris, Tokyo, Washington DC, Kyoto, and Los Angeles). The length of the route was calculated in kilometers, whereas the number of passengers was given per year in million.\nOverall, as shown in the table London is having the oldest railway system which was opened in 1863, and the longest route 394 km, compared to other cities London is having almost twice the length of the route. However, Tokyo was having the highest number of passengers which is 1927 million.\nOn the other hand, the most recent railway system opening was in 2001 in Los Angeles, where the number of passengers recorded a lower value at 50 million. In addition, the least value recorded for passengers per year was in Kyoto at 45 million.\nBoth Los angles and Kyoto are having the shortest route length which is 28 km and 11 km. But as for Paris, the route length is 199 km coming in the second longest route among the given cities.	5.5	7.0
	The graph below shows the percentage of people going to cinemas in one European country on	The bar chart illustrates the proportion of population went to the movie theaters in an European country. It divided into three periods in a varied days.\nOverall, the percentage of people going to watch films at weekend outnumbered those going in weekdays by a significant margin. While over the different three years were relatively similar in all days, there was some variance, especially on Saturday in 2005 when the percentage of individuals		

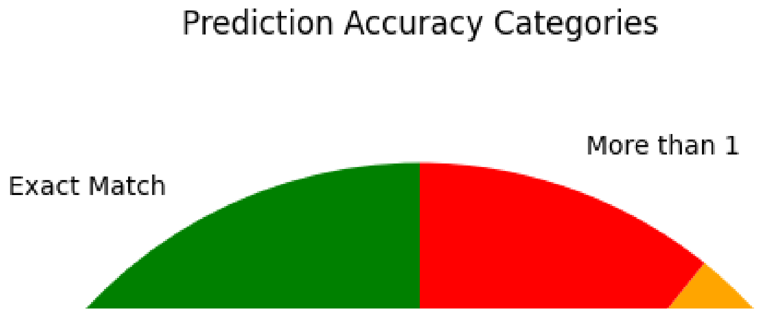
	Question	Essay	Overall	predicted_score
1256	different days.Summarize the information by selecting and reporting the main features and make comparisons where relevant.	sharply increased to almost 45 percent.\nOn weekdays, apart from Friday all periods were exactly same percentage 30%, in 2003 the proportion of people attending cinema doubled from 10% on Monday to well over 20% by the end of Thursday. Meanwhile, in 2005, over the all days it is relatively stable nearly 14% except on Tuesday where slightly increased to approximately 18%. In 2007, the trajectory was inconsistent rising from about 13% on Monday to 18% on Thursday.\nWith regards to weekend days, the records was highest on Saturday for instance in 2005 there was the most popular time for people to watch movies they were 45%, however, on Sunday the reading also high.	5.5	6.5
1218	The graphs below show the types of music albums purchased by people in Britain according to sex and age.Write a report for a university lecturer describing the information shown below.	The bar charts below illustrate the different distributions of people who bought pop, rock, and classical music based on gender and age.\nOverall, male buyers outnumbered female counterparts in all three types. At the same time, the distribution of people favouring pop music was almost identical to those buying rock music albums. However, the percentages for people of different ages fond of classical music showed a huge disparity from the other two types.\nIn detail, nearly 30% of males purchased pop or rock music albums, compared to approximately 19% of female counterparts. Among the four age groups, teenagers (16-24) and young people (25-35) were main pop and rock music supporters, with each group contributing roughly 30%. Around one quarter of people in their middle age (35-45) bought pop or rock music, whereas only one in ten people older than 45 favoured these two types of music.\nOn the other hand, regardless of their gender, classical music lovers only made half proportion as much as those who love pop and rock music. Interestingly, slightly over 20% of people older than 45 paid for classical music albums, while teenagers (16-24) and middle-aged people (25-35) accounted for no more than 3% combined in this category. Nevertheless, young adults who love classical music reached a surprising record at 16%.	7.0	7.0

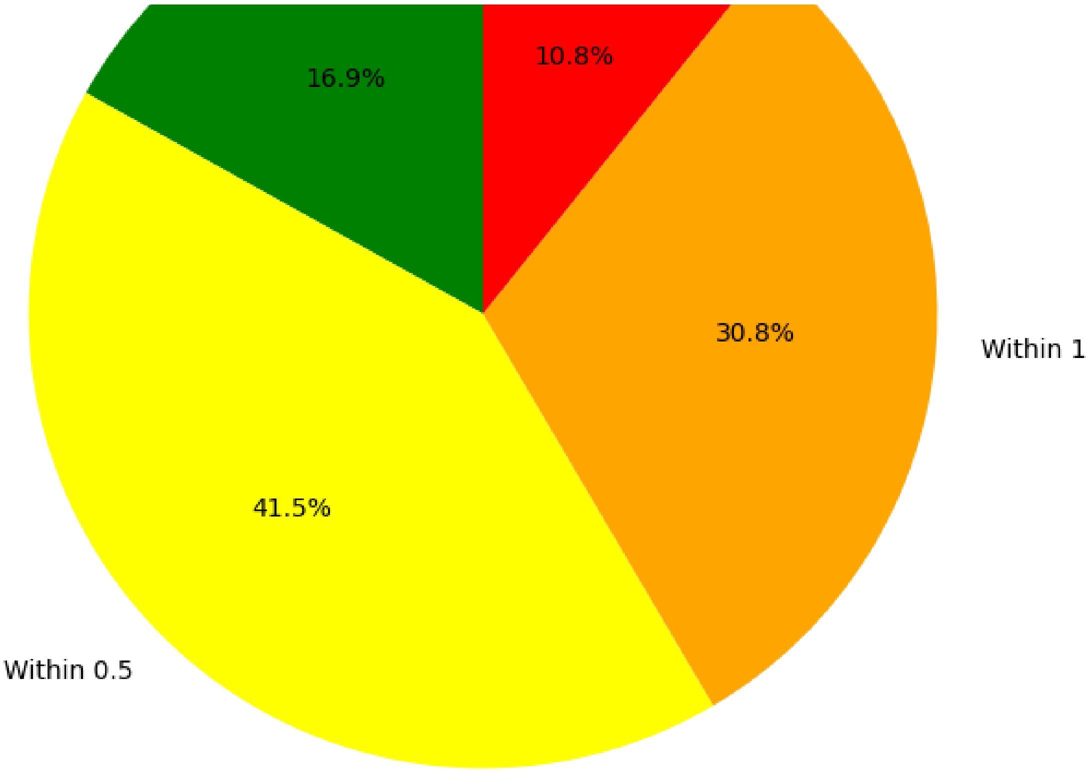
For some of these essays, the model can predict the exact score (indices 352, 1102 and 1218), while for others the predicted scores are relatively close (within 1 point) to the actual score (indices 216, 426, 886, 992, 1256). It's also notable that a few of the predictions fall outside of that 1 point window (indices 1034, 1200).

The pie chart below provides percentages for each of the rows in the test dataset to help better visualize the model's performance, with four categories: 1) Exact matches 2) Predictions within 0.5 of the actual score 3) Predictions within 1 of the actual score 4) Predictions off by more than 1 point

In [16]:

```
task_1_test['difference'] = np.abs(task_1_test['Overall'] - task_1_test['predicted_score
'])

conditions = [
    (task_1_test['difference'] == 0),
    (task_1_test['difference'] > 0) & (task_1_test['difference'] <= 0.5),
    (task_1_test['difference'] > 0.5) & (task_1_test['difference'] <= 1),
    (task_1_test['difference'] > 1)
]

categories = ['Exact Match', 'Within 0.5', 'Within 1', 'More than 1']
colors = ["green", "yellow", "orange", "red"]

task_1_test['category'] = np.select(conditions, categories)

category_counts = task_1_test['category'].value_counts().reindex(categories)

plt.figure(figsize=(8, 8))
plt.pie(category_counts, labels=category_counts.index, autopct='%1.1f%%',
        startangle=90, colors=colors)
plt.title('Prediction Accuracy Categories')
plt.show()
```





The model only achieves exact matches for 16.9% of the test data. However, it's important to note that the writing scores are inherently subjective, meaning the mistakes by 0.5 point or 1 point are not severe failures. In fact, 89.2% of the predicted scores fall within 1 point of the actual score, indicating the strength of the model in imitating IELTS grading.

Task 2 Model

We can build a similar BERT-based model for Task 2. Just like for Task 1, three Pandas DataFrames for training, validation, and testing are created from the full DataFrame with 80%-10%-10% ratio for training-validation-testing. We can again use the prepare_dataset function developed earlier to build the PyTorch Datasets.

In [17]:

```
task_2_train, task_2_temp = train_test_split(df_task_2, test_size=0.2, random_state=42)
task_2_val, task_2_test = train_test_split(task_2_temp, test_size=0.5, random_state=42)

task_2_train_dataset = prepare_dataset(task_2_train)
task_2_val_dataset = prepare_dataset(task_2_val)
task_2_test_dataset = prepare_dataset(task_2_test)
```

Since the data and the task is rather similar, we set the same hyperparameters for Task 2 as those for Task 1, only changing the output directory

In [18]:

```
model_2 = BertForSequenceClassification.from_pretrained('bert-base-uncased',
                                                        config=config)

training_args_2 = TrainingArguments(
    output_dir='./results_2',
    num_train_epochs=10,
    per_device_train_batch_size=8,
    per_device_eval_batch_size=8,
    warmup_steps=500,
    evaluation_strategy="epoch",
    save_strategy="epoch",
    learning_rate=1e-5,
    load_best_model_at_end=True,
    metric_for_best_model="mse",
    greater_is_better=False
)

trainer_2 = Trainer(
    model=model_2,
    args=training_args_2,
    train_dataset=task_2_train_dataset,
    eval_dataset=task_2_val_dataset,
    compute_metrics=compute_metrics
)

trainer_2.train()

model_path = "./results_2"
model_2.save_pretrained(model_path)
tokenizer.save_pretrained(model_path)
```

```
Some weights of BertForSequenceClassification were not initialized from the model checkpo
int at bert-base-uncased and are newly initialized: ['classifier.bias', 'classifier.weigh
t']
You should probably TRAIN this model on a down-stream task to be able to use it for predi
ctions and inference.
/Users/seyedmasihzakavi/Library/Python/3.9/lib/python/site-packages/accelerate/accelerato
r.py:432: FutureWarning: Passing the following arguments to `Accelerator` is deprecated a
nd will be removed in version 1.0 of Accelerate: dict_keys(['dispatch_batches', 'split_ba
tches', 'even_batches', 'use_seedable_sampler']). Please pass an `accelerate.DataLoaderCo
nfiguration` instead:
dataloader_config = DataLoaderConfiguration(dispatch_batches=None, split_batches=False, e
ven_batches=True, use_seedable_sampler=True)
  warnings.warn(

 10%|█         | 80/800 [02:28<17:55,  1.49s/it]
```

```
{'eval_loss': 37.36153030395508, 'eval_mse': 37.36152267456055, 'eval_runtime': 6.1311, '
eval_samples_per_second': 12.885, 'eval_steps_per_second': 1.631, 'epoch': 1.0}
```

```
 20%|██        | 160/800 [06:00<18:29,  1.73s/it]
```

```
{'eval_loss': 16.389707565307617, 'eval_mse': 16.38970184326172, 'eval_runtime': 5.6785,
'eval_samples_per_second': 13.912, 'eval_steps_per_second': 1.761, 'epoch': 2.0}
```

```
 30%|███       | 240/800 [09:07<16:40,  1.79s/it]
```

```
{'eval_loss': 3.080759048461914, 'eval_mse': 3.0807595252990723, 'eval_runtime': 5.677, '
eval_samples_per_second': 13.916, 'eval_steps_per_second': 1.761, 'epoch': 3.0}
```

```
 40%|████      | 320/800 [12:13<14:53,  1.86s/it]
```

```
{'eval_loss': 0.8504762053489685, 'eval_mse': 0.8504760265350342, 'eval_runtime': 6.3027,
'eval_samples_per_second': 12.534, 'eval_steps_per_second': 1.587, 'epoch': 4.0}
```

```
 50%|█████     | 400/800 [15:18<11:45,  1.76s/it]
```

```
{'eval_loss': 0.688987672328949, 'eval_mse': 0.6889877319335938, 'eval_runtime': 5.7165,
'eval_samples_per_second': 13.82, 'eval_steps_per_second': 1.749, 'epoch': 5.0}
```

```
 60%|██████    | 480/800 [18:21<09:51,  1.85s/it]
```

```
{'eval_loss': 1.4739162921905518, 'eval_mse': 1.47391676902771, 'eval_runtime': 6.3521, '
eval_samples_per_second': 12.437, 'eval_steps_per_second': 1.574, 'epoch': 6.0}
```

```
 62%|██████▎   | 500/800 [19:10<11:18,  2.26s/it]
```

```
{'loss': 13.7737, 'grad_norm': 74.68917083740234, 'learning_rate': 1e-05, 'epoch': 6.25}
```

```
 70%|███████   | 560/800 [21:25<06:47,  1.70s/it]
```

```
{'eval_loss': 0.742181122303009, 'eval_mse': 0.7421810030937195, 'eval_runtime': 5.7977,
'eval_samples_per_second': 13.626, 'eval_steps_per_second': 1.725, 'epoch': 7.0}
```

```
 80%|████████  | 640/800 [24:40<05:17,  1.98s/it]
```

```
{'eval_loss': 0.7955295443534851, 'eval_mse': 0.7955293655395508, 'eval_runtime': 6.3723,
'eval_samples_per_second': 12.397, 'eval_steps_per_second': 1.569, 'epoch': 8.0}
```

```
 90%|█████████ | 720/800 [28:07<02:43,  2.05s/it]
```

```
{'eval_loss': 0.48972833156585693, 'eval_mse': 0.48972833156585693, 'eval_runtime': 6.946
6, 'eval_samples_per_second': 11.372, 'eval_steps_per_second': 1.44, 'epoch': 9.0}
```

```
100%|██████████| 800/800 [31:31<00:00,  1.83s/it]
```

```
{'eval_loss': 0.8300278186798096, 'eval_mse': 0.83002769947052, 'eval_runtime': 6.7384, '
eval_samples_per_second': 11.724, 'eval_steps_per_second': 1.484, 'epoch': 10.0}
```

```
100%|██████████| 800/800 [31:35<00:00,  2.37s/it]
```

```
{'train_runtime': 1895.0298, 'train_samples_per_second': 3.346, 'train_steps_per_second':
0.422, 'train_loss': 8.716188135147094, 'epoch': 10.0}
```

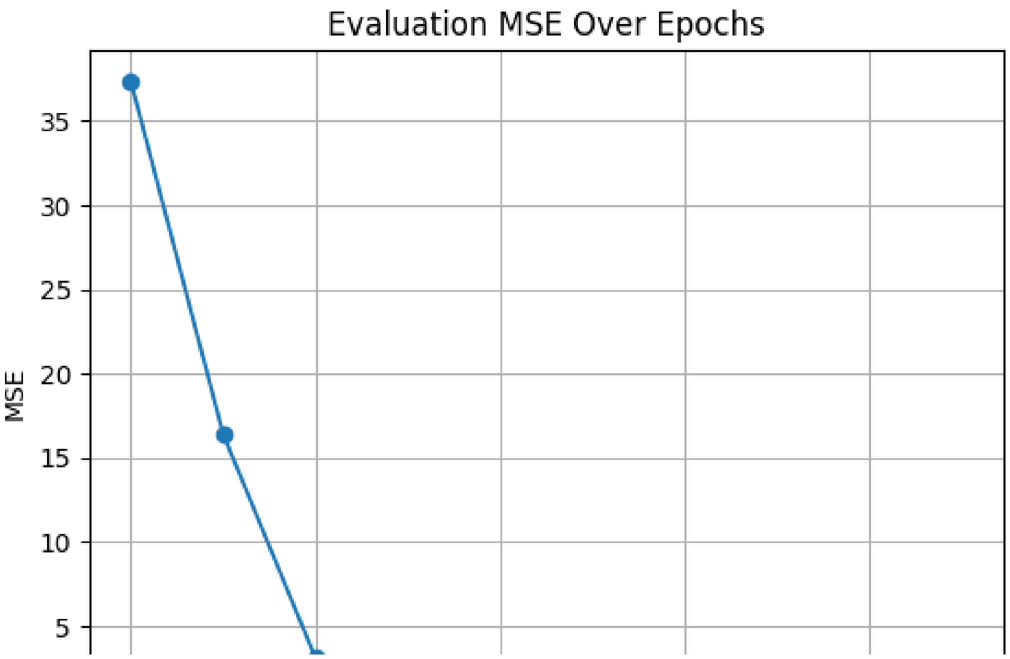
Out[18]:

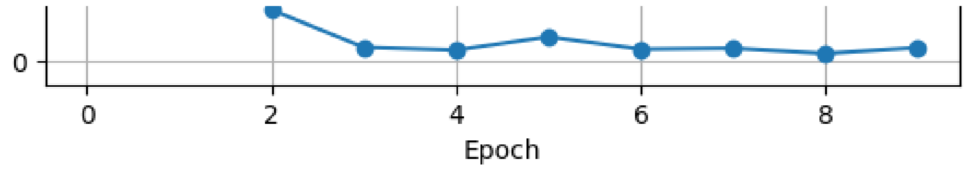
```
('./results_2/tokenizer_config.json',
 './results_2/special_tokens_map.json',
 './results_2/vocab.txt',
 './results_2/added_tokens.json',
 './results_2/tokenizer.json')
```

Like for task 1, the graph for task 2 indicates that the model has learned to grade essays for the task. This is reflected in the decline in loss, with evaluation MSE loss reaching 0.489

In [19]:

```python
with open('./results_2/checkpoint-800/trainer_state.json', 'r') as file:
    trainer_state = json.load(file)

mse_values = [log['eval_mse'] for log in trainer_state['log_history'] if 'eval_mse' in l
og]

plt.plot(mse_values, marker='o', linestyle='-')
plt.title('Evaluation MSE Over Epochs')
plt.xlabel('Epoch')
plt.ylabel('MSE')
plt.grid(True)
plt.show()
```





Just as task 1, we consider 10 actual essays to better understand the model's predictions

In [20]:

```
task_2_test_results = trainer_2.evaluate(task_2_test_dataset)
print(task_2_test_results)
```

```
100%|██████████| 10/10 [00:04<00:00,  2.18it/s]
```

```
{'eval_loss': 0.6170971393585205, 'eval_mse': 0.6170971393585205, 'eval_runtime': 6.4582,
'eval_samples_per_second': 12.387, 'eval_steps_per_second': 1.548, 'epoch': 10.0}
```

In [21]:

```
predictions_output = trainer_2.predict(task_2_test_dataset)
predictions = predictions_output.predictions

predictions = predictions.squeeze()

rounded_predictions = 0.5 * np.round(predictions / 0.5)
task_2_test['predicted_score'] = rounded_predictions
sample_df = task_2_test.sample(n=10, random_state=42)

sample_df[['Question', 'Essay', 'Overall', 'predicted_score']]
```

```
100%|██████████| 10/10 [00:04<00:00,  2.26it/s]
```

Out[21]:

	Question	Essay	Overall	predicted_score
1279	Some people believe that we cannot learn anything from the past for our life today, while others believe that history is a valuable source of information to understand human's life.Discuss both views and give your opinion.Give reasons for your answer and include any relevant examples from your own knowledge or experience.	History of human's life enriches us with pivotal informations .Some people believe no matter how valuable these informations , we often make our experiences and learn from our mistakes . However others stick with an idea that history of people can teach us alot of beneficial lessons . In my opinon , nobody can underestimate the past as it provides us many priceless wisdoms and precious informations .\nFollowing essay will discuss both sides argumaents and will give reasons for my personal opinion about topic mentioned above .\nGeneral perception propagates that we learn only from our experiences.Those people who claim this idea believe that we should not inherit the history of our ancesrties , neither their mistakes nor their success. While past had done and people cant change it , as a result of that we can't live their failure or success and we are reponsible for our life only . Nevertheless , today individuals can't think the same way our old generations had . For instance ,a person who lived centuries before had different problems , diffrent type of job , and different life style, so it's no sense to waste time digging for informations about him , therefore our choices and decisions should be based on our experiences and our knowledge only .\nHowever, in my opinon , past will provide us a perfect picture for our present and reflect the future . Whilst the development that happened in all aspects of our life had resulted from accumulated experiences of other people who lived in the past . Hence their experiences including failiures and success enrolled into our life and depict recent picture we live today . Furthermore past could leverage lessons how to avoid mistakes , for example if we read the history of some countries and understand the reasons behind revolutions and civil diputes then we can put solutions for our problems at present, and avoid falling in intesive wrestles in future. Do we have the courage to do so ?\nTo sum up, we need to learn more about the history of old people and try to figure out their problems and analyse their flaws .Hence we can solve many incoming problems we could face in future .	6.5	6.0
	Providing a national system in a country where	In modern day society, money is a driving force for nearly everyone. Most people aspire to be financially secure and to have the ability to live the life they want. However, not everyone is able to find paid employment, and for that reason in some countries around the world, governments have initiated a system where the unemployed receive a regular payment to enable them to survive. Some people believe it is an excellent idea, whilst others believe that		

	Question	Essay	Overall	predicted_score
1311	the unemployed receive a regular payment only encourages people not to seek work and puts an unreasonable strain on a country's financial resources.Discuss this statement and give your opinion.Give reasons for your answer and include any relevant examples from your knowledge or experience.You should write at least 250 words.	survive. Some people believe it is an excellent idea, whilst others believe that it is exhausting a country's financial resources.When looking at the positive aspects of this system, it can be said that it prevents individuals from having a private bankruptcy. When somebody becomes unemployed, he or she will have on-going costs such as rent for a dwelling, bills for water, telephone and electricity as well as the cost of food and several other things. Without a salary, this person will fall into debt, because they will be unable to cover these expenses. A national unemployment payment system will therefore prevent this person from losing nearly everything over time. Without such system, many people would become homeless and would potentially engage in criminal activities to survive. Therefore, one can say that this system is positive. Although stories are heard of people exploiting the system, usually the newly unemployed want to find a job fast in order to become financially independent again, meaning that the government does not have to support them for long.Although there are positive aspects of this system, one can also say that if the payment made by the government is too high, it will prevent people from looking for work actively. Occasionally, the payment people receive is higher than their potential salaries, deterring them from working and creating an on-going cost for the government. A potential solution for this could be providing free training and employment support to open up new career opportunities with higher income, which would motivate this group to re-enter the workforce.In conclusion, the unemployment benefits system has positive and negative aspects. Personally, I believe that the positive effects of the system outweigh the negative. The payment must remain on a level that helps people survive without regular salaries, but does not seem attractive to exploit over a longer period of time.	9.0	8.5
1037	Write about the following topic.It is inevitable that traditional cultures will be lost as technology develops. Technology and traditional cultures are incompatible.To what extent do you agree or disagree with this view?Give reasons for your answer and include any relevant examples from your own knowledge or experience.	It is quite evident that with the speedy development of technology, the traditional cultures in every aspect of society will soon vanish. This is because with the advancements in technology, people will start to realise that there isn't really a need for these traditional cultures to still continue. As sad un-realistic as it may sound, it is the truth. Enhanced technology and our traditional cultures are just not compatible, leaving only one to emerge victorious out of the two.\nWhat I feel about this topic is a mixture to both agree and disagree with it, because there are both plus and negative sides to this.\nTalking about the Plus sides, life will become easier. Traditional culture does not only refer to the festivities, customs, norms etc. It refers to the 'life' as a whole, which is slower than the more enhanced life brought to us by technological developments. A great example of this is the invention of the car. Before the earliest cars were invented, horses would pull carriages over long distances. The invention of the first car was a huge upgrade from these animals driven-vehicles and were beneficial too. The first and most obvious benefit was that money was saved. The horses that pulled the carriages were living things, and would die one day, causing these carriage drivers to spend loads of money on the best-quality horses. That was all replaced with the invention of the first car, making it only a one-time investment for the drivers.\nThe negative side to this is that many centuries' old traditions that hold a high level of nostalgia for communites will one day disappear. Many Indegineous african communities that have upheld their traditions for hundreds of years, are disappearing due the better job opportunities provided by technology developments.\nHece, this is something with which I have a view to both agree and disagree with.	6.5	6.5
863	Traffic and housing problems could be solved by moving large companies, factories and their employees to the countryside.Do you agree or disagree?	It is often argued that by moving large companies,factories and it employees,traffic and housing problems could be irradicated.I completely agree to this fact that these problems could be solved by taking this step.\nTraffic and housing problems has been a major problem in big cities due to the increasing population in these areas.The main contribution to the rapidly growing population are the multinational companies and the industries that have been set in the cities.These companies have a thousands of employees.As a result,if these companies are established in cities,it will lead to lot of traffic.For example,mumbai,a major city in India has uncontrollable traffic due to the vast IT sector present there.It has been very difficult for the people living in mumbai to travel even the shortest distances as it takes a lot of time to move from one place to another.The demand for housing also increases due to large number of employees from industries and huge companies which inturn leads to drastic increase in the expenses for housing.As a result,they have to compromise for small houses due to high rents.\nThe only solution for this problem is by mobilising the big companies and industries to countryside which has a lot of empty lands.By taking this step,it is both benefitial to the companies as well as the employees.The cost of establishment of the companies will be significantly low in the outskirts of the cities and the rural areas compared to the city.The employees also save a lot of time due to relatively less traffic and they also benefit to a large extent by low housing exenses.The traffic, housing problems and the pollution in the cities will also be reduced by taking this initiative.\nTo conclude,the employees, the people in the citiesand the companies as a whole could benefit by mobilising the industries and	6.5	7.5

multinational companies from the cities to the country side due to the above reasons.

	Question	Essay	Overall	predicted_score
109	The advantages of the spread of English as a global language will continue to outweigh its disadvantages. To what extent do you agree or disagree?	In today's life, English is one of the biggest common language in the world. It is often argued that the benefits of learning English as a global language will superior to its drawbacks continueously. I completely agree with this idea due to the fact that English is one of the easiest language to learn and most education system in the world involve English as one of the courses.\nEnglish should be continue to use as a global langauge because it is easy to learn. Each language in the world involves different movings of the tongue in the mouth, but some of the languages have similar tongue positions with English, such as Italian, Spanish and French. It will be difficult for people in the world to start adapting a new language and using completely different tongue positions. For example, chinese, one of the most challengeable langauge in the world, it mostly does not need the support of the tongue to touch the teeth or anything else. It is hard for people with other mother language to adapt the pronounciation. Therefore, one of the biggest advantages of English as a common language is the easyness of it.\nIn addition to the easyness of English, numorous education systems in the world include English as one of the important subjects to learn support the benefits of English that will continue to overseed its disadvantages. Since English as one of the spreading languages in the world lasts for many years. A lots of generations learn English since their childhood and English already take a huge part of their educational studies. It is hard for people to abandon this language and step out of their comfort zone. For instance, one of my best chilldhood friend learn English for almost 6 years, but since she went to Spain for two years as exchange students, she get used to the way of learn English, instead of Spanish, so she fell her grade there. It is clear that English as one of the essential courses in world, it has an important connection with the education system globally and it has a huge influence in our studies.\nIn conclusion, the easyness of learning English and the crucial role of it in the educations system support my opinion that I completely believe that the advantages of spreading English as an important language in the world will keep outweigh its flaws.	6.5	6.0
1033	Write about the following topic.It is inevitable that traditional cultures will be lost as technology develops. Technology and traditional cultures are incompatible.To what extent do you agree or disagree with this view?Give reasons for your answer and include any relevant examples from your own knowledge or experience.	Many people hold the view that traditional culture and technological development cannot coexist as the latter catch on, the former will be bound to disappear. I totally disagree with this because both can be implemented jointly by putting technology at the service of tradition and through pilot projects.\nTo begin with, technological progress, which is making major strides in our daily lives, is the means by which local mores, custom and traditions could be brought back to life, proving they are compatible. In fact, by tapping technology to make tradition far more attractive, considerably more audience will be engaged into such activities. Bologna is a case in point, since, thanks to digital machines and drones which have supported housewives into the making of , the traditional celebration of “making pasta day ” has recently being turned into a successful events, which draws people from country as far afield as China.\nAnother example of possible coexistence is the possibility to heighten awareness of people costumes through the creation of pilot projects which bring both habits and technology to a new whole level. This can be put into practice by setting a direct connection to cultural heritage of a specific area of the world to those who do not have access easily. To this end, many primary schools in New York have developed a project in which, owing to the ultimate high speed connection, pupils who come from abroad can learn ancient craft directly from a master located in their country of origin.\nTo conclude, I firmly believe that traditional culture and technology development are widely compatible as the technology can positively beef up the presence of cultural traditions nowadays, so as to keep its memory Alive. In addition, pioneering projects can be implemented to underpin cultural awareness amongst youngsters.	7.0	7.0
	Write about the following topic.Most people have forgotten the meaning behind traditional or religious festivals; during festival periods, people nowadays only want to enjoy	Nowardays, it widely believed that most people have forgotten the main purpose of some traditions and religious celebrations, but during those festival periods they just prefer to enjoy themselves and have a fun from holiday. In this essay, I will discuss two opposite attitudes of modern people to some traditional celebrations and give example and some advice to the issue.\nOn the one hand, it seems that the majority of people does not care about the reason of holidays, they are ready to enjoy them without any concrete reasons. First of all, that happens usually because of the fact that many people overwork and do not live in their work-life balance, so they are very glad to have additional day-off. The second explanation to the issue can be that people follow some traditions from ancient times, so some of those traditions became as a routines, that the basic meaning could be forgotten during the times. But nations continue folloving some customs just because they got used to them, even if they do not remember the main purpose of the celebration. The third reason to that can be that young generation do not understand some of their ancestors traditions, so they just accept some festivals from their predecessors as days when they are able to make their		

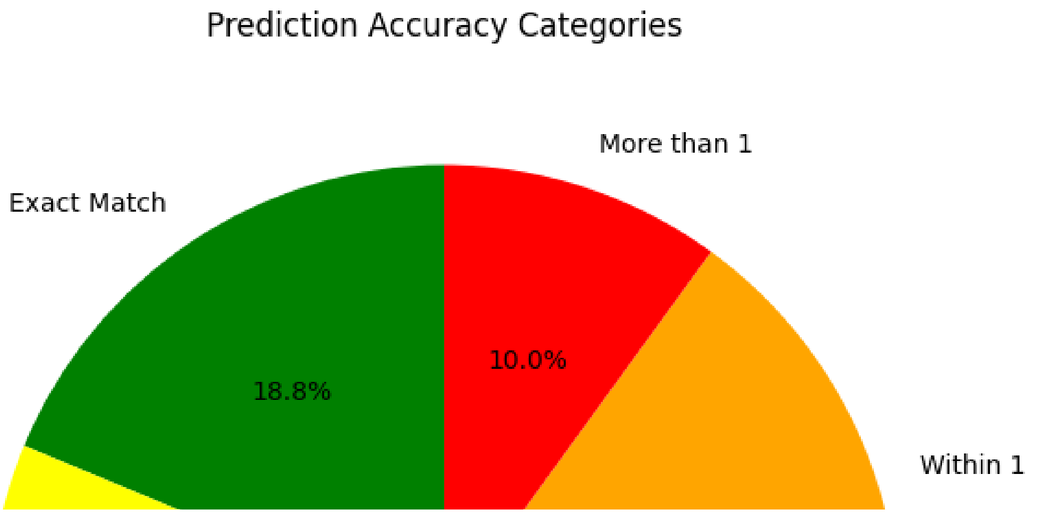
	Question	Essay	Overall	predicted_score
1047	themselves.To what extent do you agree or disagree with this opinion?Give reasons for your answer and include any relevant examples from your own knowledge or experience.	lives brighter and enjoy themselves.\nOn the other hand, there are exist people who know the basics of festivals and familiar with the hystory of traditions in their countries. Such kind of people usually consern about their traditions and pay more attantions to some specific atributes of festivals, which can be symbolise health, wealth, success, or maybe femily hapiness, et cetera. In my opinion, it would be beneficial if the main purposes of some, especially ancient, traditions, to be explained to others. First of all, people would become familiar with some old customs and maybe partly could understand the reasons why their ancestors used some kind of colors and accesuars during that holidays. Additionally, people would know better the histories of their countries and share those knowledge with younger generations. So, the youth could not only enjoy themselves during celebrations, but also become more intelligent about traditions in their homelands.\nTo sum it up, it is widely spread attitude on holiday organisation as to only having entertainment, dancing, and having wonderful moods during those days. However, if the history of festivals and ancient traditions is spread properly, people become more wisdom about the main reasons of those celebrations in their countries.	6.0	6.0
1291	The end of the world’s fossil fuel resources and the subsequent changes will be a positive development in society. To what extent do you agree or disagree with this statement?Give reasons for your answer and include any relevant examples from your knowledge or experience.You should write at least 250 words.	Today’s society overwhelmingly depends on products made from fossil fuels. Most people would initially think about fuel for transport, but of course, fossil fuels are part of most of the fabric of life around us, from the plastics on our furniture to the pavements that we walk on. Changing our reliance on these resources can have various consequences; however, I believe that overall the society will benefit from finding viable alternatives for them.Depending on whether eco-friendly alternatives are developed, the positive effect on transport and the reduction of the associated pollution will be significant. Exhaust fumes from motor vehicles, planes and ships decrease air quality and contribute to global warming and the greenhouse effect. People’s health and the health of the planet will surely improve without these industries relying on fossil fuels. It could be that a type of transport crisis could develop as a result if society can no longer provide the power to transport so many individuals to where they want to go. This, however, does not have to be a bad thing, as it could force society to use alternative solutions, such as transport sharing and public transport development. The change in mind set has already begun. Some countries have already committed to having fossil fuel free transport and more cities are banning private vehicles from the centre of cities.Another positive development would be that more time and money would have to be put into developing renewable sources of energy, such as wind power, solar power, hydropower and tidal power. Technologies would improve and be more efficient and again, the world would use fewer sources of energy that cause pollution. This development can also be seen today, with many governments committing to proportions of their power needs coming from renewable resources.I, therefore, strongly agree with the statement that the changes from reducing reliance on fossil fuels will be positive. The only caveat is that society will need to have developed alternative sources of power rather soon, in order to avoid an energy crisis.	9.0	8.5
601	Write about the following topic.Group or team activities can teach more important skills for life than those activities which are done alone.Do you agree or disagree?Give reasons for your answer and include any relevant examples from your own knowledge or experience.	Group work is believed to foster more life skills than activities done solo. Although doing things alone is beneficial as it strengthens decision making skills, working in teams can educate more valuable life lessons as it teaches teamworking and negotiation skills necessary for living in a community.\nFirstly, working in a group helps build teamworking skills. In a team, in order to be successful in doing peojects, members need to help one another in achieving mutual goals. Therefore, through these steps, team members will be trained on assisting others to reach the goal, although they may feel more competent undergoing a project on their own. Similarly, in real societies, people living in the same community cannot be careless to their neighbors and, instead, have to understand each other. Because of this understanding, they can tolerate one another.\nStill, it can be argued that working collectively nurtures people to follow the crowd and hence, they cannot sharpen decision making skills. For this point, it is valid that a lone wolf has a clearer vision in its life direction than a pack of wolves that walk behind a leader. Still, a single leader may lead to a wrong path that ends up to a disaster for all other wolves. Thus, in order to make a right decision effectively, it is required that people work together since they can debate and negotiate what is appropriate for the team which, in turn, beneficial for each individual. Likewise, in a country, no single prime minister can dictate the country's direction. Nevertheless, opinions gathered from the parliament are necessary.\nIn conclusion, team projects are more useful in culminating skills crucial to live in a society. Teamworking skills gained from working together will lead to an understanding of other people not only in the team but in the same society. Moreover, people will be trained on negotiation and compromisation which yield better success than decisions done alone by a single person.	7.0	8.0
		Nowadays, many businesses have to compete each other in order to survive from the rapidly changing world, leading to the high amount of time that		

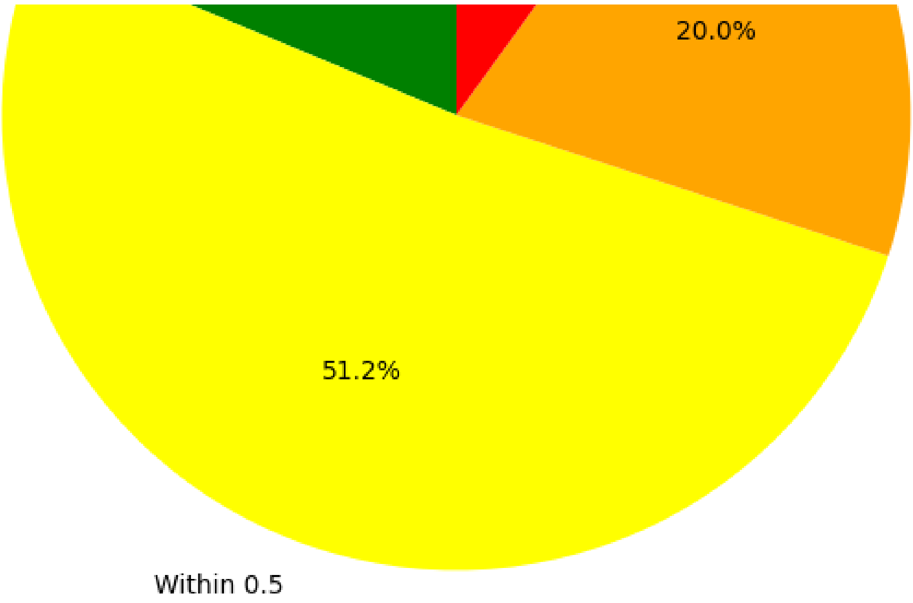
	Question	Essay	Overall	predicted_score
681	In some countries people spend long hours at work. Why does this happen? Is it positive or negative development	employees spend on working. There are also other reasons why the employees work for a long period which I will discuss in this essay. In terms of consequences, I personally believe that spending most of our time at work is the negative development.\nIt is the fact that people have various rationales to work for long hours. A classic example is the dynamic working condition. There are many companies that provide services such as consultation to customers, meaning that if their service does not completely satisfy the clients' need, such client may decide to use the service from other competitors. Hence, the employees are required to dedicate their time to deliver the best service to the clients. Another common criticism is a bright career path. When the employees work for a long period, they can carefully check their works and, as a result, delivery the works with less mistakes. Doing this way could impress their manager and they are likely to get promotion in return.\nWith a careful consideration regarding the consequence of working for long hours, there are many vital drawbacks. An obvious effect is the increase in illness. People who work very hard have to spend a lot of their time at work. Therefore, they may not have enough time to do exercises, weakening the immune system of their body. Furthermore, with the rise in the amount of working hours, people tend to lack social skills. This is because they may not be able to manage time to join social events or meet their friends, resulting in the less ability to adapt to new environments.\nIn conclusion, it is undeniable that people work for long hours because of the requirement in business and personal career goal which, in my opinion, leads to negative development for themselves. This is because this behavior negatively affects their health condition and social competency which are the two most important aspects in our life.	7.0	8.5

Compared to task 1, the model's predictions seem to be closer to the actual grades for task 2. Out of the 10 randomly sampled essays, only one shows a severe (more than 1 point) mistake in predicting the grade. The pie chart for all the test data can provide a better means for comparing the two models

In [22]:

```
task_2_test['difference'] = np.abs(task_2_test['Overall'] - task_2_test['predicted_score
'])

conditions = [
    (task_2_test['difference'] == 0),
    (task_2_test['difference'] > 0) & (task_2_test['difference'] <= 0.5),
    (task_2_test['difference'] > 0.5) & (task_2_test['difference'] <= 1),
    (task_2_test['difference'] > 1)
]

categories = ['Exact Match', 'Within 0.5', 'Within 1', 'More than 1']
colors = ["green", "yellow", "orange", "red"]

task_2_test['category'] = np.select(conditions, categories)

category_counts = task_2_test['category'].value_counts().reindex(categories)

plt.figure(figsize=(8, 8))
plt.pie(category_counts, labels=category_counts.index, autopct='%1.1f%%',
        startangle=90, colors=colors)
plt.title('Prediction Accuracy Categories')
plt.show()
```





As expected, the model for task 2 is better than the one for task 1. There are more exact matches and the number of severe mistakes has also marginally decreased. More notably, 70% of the preedictions are within 0.5 of the actual score, far more than the 58.4% for task 1

Discussion

The Model's superior performance on task 2 can be explained by more training data for task 2 and the longer essays allowing for more consistent grading. Another key factor for understanding the better performance is the absence of the visuals for task 1. The model has no means of realistically verifying the essay's understanding of the chart and graph since those are not provided to the model. Since task achievement for task 1 is providing a clear description of the visual, this severely limits the model's ability to make accurate predictions. A better model would condition the grading on a representation of the visual. However, this would require a multimodal transformer and more computational resources.

It's also crucial to note that the amount of data for both of these tasks is very limited due to privacy reasons. Even with only 1435 tasks provided in total, BERT achieved close predictions for 90% of the essays for both tasks. With more data and better diversity in the data, its plausible that BERT could achieve even higher accuracy in grading.